Improving headline school performance measures

Multi-year averages should be used in school league tables if they return in 2021.

Loic Menzies and John Jerrim

If performance tables return for the 2021 Year 6 and Year 11 cohorts, the Department for Education will face important questions regarding whether they can provide reliable and valid information about the "performance" of a school. This is because of the substantial amount of time children currently in Year 5 and Year 10 will have spent out of school in the recent past, with their learning over this period to a great extent outside of the control of schools.

Head teachers have already warned that it would be "unfair" to publish league tables based on headline exam results when *"The education of the children taking these assessments has already been disrupted by the coronavirus lockdown, and it is likely that there will be further disruption next academic year [1]".*

We agree that it would be unfair to publish headline school performance measures based solely on the 2021 exams series. A better approach would be to take an average across at least two years, from the 2019 (pre-Covid) and 2021 cohorts.

We believe the current situation therefore strengthens the existing case for school performance measures based on multi-year averages.

In this short discussion paper we summarise our evolving thinking in relation to our proposed approach, drawing on a recent roundtable attended by leading academics and government officials.

[1] The Observer, (2020) https://www.theguardian.com/education/2020/jul/05/teachers-urge-suspension-of-english-school-league-tables-in-2021

1. Upcoming Challenges

A number of challenges will arise if school performance tables return in 2021 after a Covid-induced hiatus. None of these are entirely new, but all will have been exacerbated by the pandemic:

1. The influence of school and pupil context on the performance measure.
As research by George Leckie and others has shown, school performance measures, are not, in reality, pure measures of ***school*** effectiveness. They could potentially be improved for this purpose if they controlled for various aspects of pupil background – although this would not resolve all the issues[2].

The Covid crisis and associated school closures are likely to amplify the impact of contextual differences and therefore diminish the extent to which league tables can claim to offer true measures of *school* performance in their current form[3,4]. This might create a sense in coming years that league tables are unfairly, and unequally punishing schools for the way Covid has impacted on pupils.

2. Instability
School performance measures are always, to some extent, volatile and this problem is particularly marked in small schools.

A schools' performance moves around from year to year due to a combination of marking unreliability, teacher turnover, cohort effects and external shifts that impact on different schools differently depending on their characteristics. For example, the impact of a decision to include or exclude a qualification from league table measures will depend on the extent to which the school previously used that qualification. Meanwhile, schools are increasingly moving between MATs and being given new URNs when they 'shut', making it harder to look at performance over time.

Ongoing school disruption due to Covid-19 will result in yet another source of instability.

3. Missing or discontinuous data
Performance tables have been affected by missing data a number of times in the past. In 2010 a quarter of schools boycotted the government's SATs and this not only impacted on primary school results that year, but also on progress measures at KS4, five years later. This caused difficulties during previous attempts to construct trend data for MATs using multi-year data.

Meanwhile, although the original intention when introducing Progress 8 was to include multi-year measures[5], this has been delayed due to GCSE reforms, including the shift to number grades, making it difficult to produce comparable data. Further challenges are anticipated with the introduction of T levels.

---

[2] Leckie, G., & Goldstein, H. (2019). The importance of adjusting for pupil background in school value-added models: A study of Progress 8 and school accountability in England. British Educational Research Journal, 45(3), 518–537. https://doi.org/10.1002/berj.3511
[3] The Sutton Trust (2020), 'COVID-19 IMPACTS: SCHOOL SHUTDOWN'
[4] Institute for Fiscal Studies (2020), 'Learning during the lockdown: real-time data on children's experiences during home learning'
[5] Department for Education (2013) Reforming the accountability system for secondary schools Government response to the February to May 2013 consultation on secondary school accountability https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/249893/Consultation_response_Secondary_School_Accountability_Consultation_14-Oct-13_v3.pdf

Missing data in 2020 will not only create a gap in time series data, it will also have long term repercussions in terms of calculating progress measures (e.g. in 2025, when current Year 6 pupils sit their GCSEs).

4. Communication to different audiences
Performance tables are designed to balance a number of different purposes (see below) and to cater for the needs of multiple different audiences.

What might be appropriate to communicate the most robust possible information to expert audiences is not the same as what is appropriate for parents and schools. Lay audiences have limited understanding of how the progress 8 measure is calculated and a balance has therefore had to be struck between building a sophisticated P8 metric, and one that is sufficiently intuitive for a range of audiences to navigate. Furthermore, it not always the DfE that decides what measures are prioritised since reporting on exam results is a big focus for newspapers and what they chose to focus on tends to be what ends up being valued.

In the absence of exam data for 2020, Ofqual and the government have had the unenviable task of finding a way of awarding grades that is fit for purpose, but it is clear that understanding how grades are being awarded is causing considerable confusion and anxiety to young people, parents and schools[6].

We have long felt that drawing on more than one year's data to construct multi-year averages would mitigate some of these long-running challenges, as well as various other dysfunctional aspects of the current system[7]. It is not clear though whether recent events make our approach more needed than ever, or whether the potential additional benefits are outweighed by the heightened difficulties associated with a year's missing data.

2. Are performance tables telling us what we need them to and would multi-year averages do that better?

Performance tables exist to do a number of distinct things and constantly trade these off in a series of unsatisfactory compromises[8]. The question of whether multi-year averages would be a better approach to a headline measure cannot escape the question of purpose.

School performance measures might be said to exist:

1. To show how well a school is doing
    a) so that parents can make informed choices
    b) so that schools can be held accountable.
2. To set expectations and shape behaviour across the system.

Their explicit purpose appears to be either 1a or b, but the importance of 2 cannot be ignored and this is clearly, from government's perspective, the main blocker to more contextual measures.

[6] Yeeles et al., 2020, Assessing the early impact of school and college closures on students in England https://cfey.org/reports/2020/06/assessing-the-early-impact-of-school-and-college-closures-on-students-in-england/
[7] Jerrim, J., Menzies, L. (2019), https://cfey.org/2019/11/accountability-reformed-the-case-for-multi-year-measures/
[8] Millard, W., Small, I., & Menzies, L. (2017). Testing the Water: How assessment can underpin, not undermine, great teaching.

In terms of 1a versus 1b, there has been a shift away from 1b, because the trigger for intervention has moved towards inspection and away from quantitative measures and floor targets.

Thus, the critical question is whether multi-year averages would do a better job of helping parents to make informed choices, whilst avoiding unforeseen problems for 1b and 2.

Performance measures and school choice
It is questionable whether school performance tables can ever be a good way of helping parents make good school choices (and of course, whether school choice is in itself desirable or feasible).

> *"Relying on league tables to inform school choice leads to highly misleading judgments since these tables ignore the uncertainty that arises from predicting schools' future performance based on their past performance. We have shown that, when taking account of this uncertainty, the comparison of schools becomes so imprecise that, at best, only a handful of schools can be separated from the average school or from one another with an acceptable degree of precision. This implies that publishing league tables to inform parental choice of school is a meaningless exercise, as parents are using a tool which is not fit for that purpose[9]."*

On the other hand Rebecca Allen and Simon Burgess have shown that, using performance tables, parents make 'the right' school choice twice as often as they make the wrong choice[10].

Either way, from a pragmatic point of view, it is clear that the current policy view from government is that school choice is a priority, and that performance tables should help parents to make these choices.

It could be argued that knowing 'how well pupils tend to do', drawing on a larger sample (particularly in small primary schools) is more useful to parents than just knowing 'what happened last year'.

On the other hand, perhaps parents would prefer to get a picture of what is happening *now*, drawing on the most up to date information. In this case, multi-year measures may be problematic given that current measures are already lagging measures, because results in 2019 show (amongst other things) how well a school was doing between 2014 and 2019. Moving to three year averages would shift this to 'between 2012 and 2019'. Schools that have underperformed in the past but that are now on a rapid journey of improvement could be particularly disadvantaged by this, and an albatross around a school's neck could make recruiting school leaders in challenging school even harder.

This, in our view, boils down to what weights are used. It would, for instance, be possible to still give the most recent year of data the most prominence (e.g. 50% of the final score) while not entirely discounting valuable information on past performance, as the current single-year approach does (e.g. results from the preceding two years could be weighted at 25%). Variable weightings would not be unprecedented since certain subjects (e.g. maths and English) already receive more weight than others under Progress Eight.

[9] Leckie, G., & Goldstein, H. (2009). Are League Tables any use for choosing Schools? Research in Public Policy, Summer, 6–9.
[10] Allen, R., Burgess S., (2013) Evaluating the provision of school performance information for school choice https://www.sciencedirect.com/science/article/abs/pii/S0272775713000289

Emerging Question 1: What is the best type of information to provide to parents to help them make informed choices and in what form, considering both what they prefer, and what leads to 'optimal choices' according to an externally set criteria (like the one used by Allen and Burgess)?

Usability

Government wants parents to find the data it publishes simple to use, but we are not aware of evidence regarding what information parents currently use and how helpful they find this.

There are important questions about the usability of a school performance measure based on several year's data, but it is important to distinguish between complex computation and complex presentation.

The shift from 5 A*-C including English and Maths to P8 involved a considerable jump in the complexity of the calculations underpinning the headline measure. However it is not necessarily the case that the change made the figures a great deal harder to use, particularly once it is presented as "well above/below average" as it is on the current "Compare School Performance" website[11].

Indeed, small-scale, focus group-based research by the DfE suggests that parents focus on the 'banding' element of current measures, but in some cases defer to a simple percentage achieving English and Maths when this seems at odds with the banding. Ofsted's quintile measures have also proved popular with inspectors and these draw on multiple year's data.

Ultimately, defining the ingredients of a measure is therefore not the same as defining how it is presented, and it is likely to be the latter that matters more when it comes to usability for parents.

Emerging Question 2: To what extent would shifting the 'ingredients' of a measure worry parents, if the presentation remained relatively unchanged?

Emerging Question 3: How do parents read school performance information and what do they value? Could randomised trials and eye-tracking with screen testing help to understand this?

3. What are the key technical considerations to take into account when building a multi-year average?

Reliability

A reliable measure would remove noise and secure a good 'school quality' signal. However it should not be assumed that using more than one year's data is *necessarily* more reliable. There is also a question about the role of sampling errors versus measurement errors and which are the primary cause of unreliability. This is particularly important given that when it comes to performance measures, error is introduced at both baseline (e.g. Key Stage Two) and the outcome (e.g. Key Stage 4 level),

[11] https://www.compare-school-performance.service.gov.uk/

exacerbating unreliability. This is true for both primary school (e.g. Key Stage 1 to Key Stage 2 value added) and secondary school (e.g. Progress 8) measures.

Emerging Question 4: Can we test the reliability of different measures and assess impact of multi-year averages on reliability?

Number of years and weighting

There is a question as to the optimal number of years to include. This is a parallel question to weighting, since, in effect using one year's data means 100% weighting one year and zero weighting the other years. There is therefore a linked question of whether to weight, say, year four at 0% or to draw on that year too. The best approach will depend on the time series properties of school performance data.

Emerging Question 5: How many years of data should the multi-year averages be based upon? How should the different years be weighted?

Weightings could potentially change from year to year, for example reducing the weighting where there is disruption or reason to doubt the results. For instance, results from a year such as 2020 could be 'down-weighted' and others (e.g. 2019 and 2021) 'up-weighted'. Given that it could create difficulties for policy makers if weightings become contested every year it may be preferable to the DfE to keep weightings stable and only adjust in exceptional circumstances like 2020. Alternatively precision weighting taking into account cohort size could be an option though this would impact on confidence intervals.

Missing data

Any future-proof headline measure needs to cope with missing data, as this has been an issue in the past and will be in the future if action is not taken in response to the crunch point we have reached in 2020.

Under the current system the cancellation of Key Stage 2 tests in 2020 means there will be no baseline measure available to calculate Progress 8 scores in 2025. Similarly, there could also be issues due to an unreliable baseline in 2026, given the reduced time that current Year 5s will have spent in school.

When the SATs boycott took place, teacher assessments were used instead of exam results, but this will not be an option now that teacher assessments at KS2 have been reduced to 'meeting standards' or not. The government has also pledged not to use this year's teacher awarded grades in measures of school performance so they could not be used in future multi-year measures without reneging on this commitment.

In 2025, the most likely approach the Department for Education will take is therefore to either impute a baseline or to directly use another prior achievement measure (e.g. Key Stage 1 scores). The DfE therefore needs to start testing the extent to which different models can predict performance, using context, KS1 and EYFS data[12]. It is not unlikely that a fairly good measure can be constructed, though research underway at the

[12] See, for example, https://ffteducationdatalab.org.uk/2020/05/could-progress-8-still-be-calculated-in-2025/

moment demonstrates the limitations of such an approach[13] and it may still not be very palatable to schools to be told that they will be held account based on predicted baselines. It is also worth considering the difference between school, and pupil-level missing data.

Rather than publishing a headline school performance measure based solely upon one year's questionable data in 2025, a much better approach would therefore be to combine this with the 'better quality' information on pupil achievement at a school that is available from previous years using a multi-year average. Uncertainty in the 2025 Progress 8 data could then be reflected in a reduced weighting with the precise weighting decided in consultation between government, experts, teachers and union representatives. This would help strike the right balance between recognising school improvement and disadvantaging schools on the basis of tenuous data.

Emerging question 6 What is the most accurate model for calculating a predicted pupil baseline (or should a separate baseline test be administered next year?)

Emerging question 7: What are schools' attitudes to different approaches to calculating a baseline to be used in future progress measures?

Changing distributions

Progress 8 measures are not directly comparable across academic years. This is largely a function of the measure not truly having an absolute scale. Questions have therefore been raised about how the data could be combined across academic years and whether the distribution will need to be standardised. This challenge came up in the past when calculating MAT performance trends.

This would seem to be a relatively straightforward problem to solve, and would not involve a huge amount of complexity. Specifically, each year Progress 8 measures should be converted into a z-score, with a mean (national average) of zero and standard deviation of one. A (standardised) value of one would then, for instance, have the same meaning with respect to the *relative* performance of schools compared to the national average across academic years. These standardised Progress 8 scores could then be averaged for each school across the academic years included in the multi-year average calculation.

Emerging question 8: How should changing distributions be dealt with in a multi-year measure?

[13] Leckie, G., Prior, L., Goldstein, H. (2019). The implications of Labour's plan to scrap Key Stage 2 tests for Progress 8 and secondary school accountability in England. arXiv: 1911.06884 [stat AP]. https://arxiv.org/abs/1911.06884

4. Conclusion

Our proposal for multi-year averages enjoys widespread support amongst school leaders, unions and many policy experts.

Our roundtable demonstrated that the approach is indeed technically feasible and, whilst disruption to exams in 2020 may initially have seemed to present additional challenges, it may in fact demonstrate the urgent need for change, and the potential to mitigate several longstanding shortcomings of England's accountability system.

We look forward to convening a second roundtable in September 2020 to explore sector leaders and policy makers' appetite for change.

5. Acknowledgements

We would like to thank all the participants in our roundtable. The views in this paper are not necessarily theirs but we are grateful to them for their insight and critique.

The contributors were:
Jon Andrews (Education Policy Institute), Simon Burgess (Bristol University), George Leckie (Bristol University), Josh McGrane (Oxford University), David Robinson (Education Policy Institute), Alex Sutherland (Behavioural Insights Team), Dave Thomson (FFT Education Datalab), representative from Pearson, two DfE Officials and an Ofsted Official.